DOCUMENT RESUME

ED 360 830 FL 021 382

AUTHOR Chalhoub-Deville, Micheline
TITLE Performance Assessment and the Components of the Oral Construct across Different Tasks and Rater Groups.
PUB DATE [93]
NOTE 17p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Arabic; College Students; Comparative Analysis; Higher Education; *Interrater Reliability; Interviews; Language Teachers; *Language Tests; Native Speakers; *Oral Language; Reading Aloud to Others; Second Language Instruction; Second Language Learning; *Second Languages; Tape Recordings; *Test Format; Testing; Uncommonly Taught Languages; Verbal Tests

ABSTRACT

This study investigated whether different groups of native speakers assess second language learners' language skills differently for three elicitation techniques. Subjects were six learners of college-level Arabic as a second language, tape-recorded performing three tasks: participating in a modified oral proficiency interview, narrating a picture depicting a story, and reading a text aloud. The recordings were rated by three groups: 15 native Arabic-speakers teaching in the United States, 31 non-teaching native Arabic-speakers living in the United States, and 36 non-teaching native Arabic-speakers living in Lebanon. Ratings were given both holistically and on a nine-point scale of proficiency. Three response dimensions were assessed specifically: grammar/pronunciation; creativity in presenting information; and amount of detail provided. Results indicated variability of performance across tasks as well as between individuals. In sum, it was found that oral ability, tasks, and raters all affected students' scores. Further analysis of the effects of different tasks and of different raters on assessment of second-language performance is recommended. A 23-item bibliography and analysis data are appended. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 360 830

Performance Assessment and the Components of the Oral Construct Across Different Tasks and Rater Groups

Micheline Chalhoub-Deville

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Micheline Chalhoub-Deville

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL 021382

BEST COPY AVAILABLE

Performance Assessment and the Components of the Oral Construct Across Different Tasks and Rater Groups

Micheline Chalhoub-Deville

The Ohio State University
287 Arps Hall
1945 N. High Street
Columbus, Ohio 43210
e-mail: MCHALHOU@MAGNUS.ACS.OHIO-STATE.EDU

I Theoretical Background

Second language (L2) oral testing increasingly call for more performance-based tests. Performance-based tests require students to produce complex responses integrating various skills and knowledge; and to apply their target language skills to life-like situations. Performance assessment, however, presents a new set of problems, central to all being the issue of validating scores obtained from these tests. What do total scores obtained from L2 oral tests mean? Validity researchers concur that the primary "purpose of construct validity is to justify a particular interpretation of a test score by examining the behavior that the test score summarizes" (Moss, 1992: 233). The fundamental requirement for establishing the validity of L2 oral test scores, therefore, is defining the L2 oral construct.

According to Messick (1993), the purpose in construct validation studies is "for construct-relevant variance to cumulate capitalizing on the positive features of each response format [method] ..., while biases attendant upon the smaller construct-irrelevant variance would not cumulate as much" (p. 62). In construct validation researchers need, therefore, to specify the attributes of the construct and minimize factors that confound test score interpretation. In L2 oral language testing research, two principal factors influence and potentially confound the scores reflecting the learners' oral construct: the test method and the rater.

A potential source of confounding and irrelevant variance in language test scores is the task or test method. It is well known that test methods and different elicitation tasks influence results differentially, limiting interpretation of constructs (Bachman, 1990). Second language acquisition (SLA) researchers document that diverse elicitation tasks produce variations in speech products (Ellis, 1985, 87; Larsen-Freeman, 1991; Tarone, 1983, 1989). Tarone (1989) writes that "[t]here can be no doubt now that the linguistic forms produced by second language learners vary markedly as those learners move from one situation to another and one task to another" (p. 13).

Furthermore, L2 testing researchers concur that a test score is influenced to a large degree by the method used to measure the trait (Bachman, 1990; Bachman and Palmer, 1981; Clifford 1981; Shohamy, 1983, 1984). Clifford (1981), reviewing multitrait-multimethod studies, argues that researchers fail to provide evidence of construct validity for the traits the test purports to measure because they fail to account for the effects of test method on test scores. In summary, both SLA and L2 testing researchers document the effect of tasks and methods on learners' test scores. In investigating a construct, information from diverse

methods is needed in order to arrive at a richer and more dynamic picture of the construct.

Although much thought should be given to the selection of tasks used to elicit language samples, scoring procedures are also critical. Test scoring is another potential source of irrelevant variance that affects score use and interpretation. L2 oral tests are often human-scored, meaning that raters assign scores. Therefore, the influence of the rater on scores obtained is a potential source of error that may influence learners' scores of L2 oral ability.

Trained teachers are usually asked to assess the oral ability of foreign language learners. Many L2 learners, however, wish to communicate with native speakers, who most probably do not share the teacher raters' professional training. Teacher training may influence teachers' assessment and render their judgement different from non-teaching native raters (Engber, 1987; Shohamy, Gordon, and Kraemer 1992). Rater groups, therefore, may differ in judging learners' L2 oral ability depending on the set of criteria with which they operate.

Research shows that trained teachers and non-teaching native speakers differ in their assessment of L2 speakers' oral ability (Barnwell, 1989; Chastain, 1980). A number of studies have also documented a difference between native speakers residing in the learners' community and those living in the target community (Barnwell, 1989; Galloway, 1980). Native speakers residing in the learner's community may have become accustomed to dealing with non-native speakers, both linguistically and culturally, which may influence their ability to evaluate the L2 learner's oral ability. Consequently, assessment of learners' oral ability obtained from naive native speakers who reside in the learner's target language community may provide another perspective of the L2 oral construct. In summary, both tasks and raters warrant attention as to their potential influence on the validity of the oral construct.

II Purpose

The present study was designed to address the following research questions:

1. What are the dimensions that underlie three elicitation techniques: oral interview, narration, and read-aloud?
2. Do three groups of native speakers of Arabic--teachers of Arabic as a foreign language (AFL) in the U.S., non-teaching Arabs residing in the U.S. for at least one year, and non-teaching Lebanese--assign the same weights to the dimensions that underlie the three elicitation tasks?

III Methodology

1 Speech Samples

Six subjects provided the speech data for this study. The six subjects were AFL students. They all had completed at least four quarters of college-level study of modern standard Arabic (MSA), and were enrolled at the time of the experiment in an intermediate level Arabic language class. Two of the subjects were males and four were females. Their ages ranged from 23 to 33.

Subjects were audio-taped performing three tasks: a modified oral proficiency interview, a narration of pictures depicting a story, and a read-aloud of a text. These tasks were chosen to elicit a variety of speech products, in order to tap a wide range of the subjects' L2 oral language abilities. The learners were interviewed by a certified oral proficiency tester of Arabic. The modified oral interview, which lasted approximately 10 minutes, was employed to elicit the subjects' most spontaneous speech (Omaggio, 1986). A two-minute segment, judged as representative of a subject's performance, was included on the tape used for rating. The selection of the segments was validated with the interviewer. These segments were selected from the middle portions of the interviews. Thus, speech from the warm-up and wind-up stages, which are meant to make subjects feel comfortable and permit subjects to leave the interview situation with a sense of accomplishment was not included.

The narration task was based on a sequence of six cartoon drawings depicting a story of an encounter between a female and a male bicycling in a park. This task was considered relatively more controlled than the interview, but still effective in providing subjects with the "opportunity for personal expression and interpretation" (Underhill, 1987: 67).

The third task involved reading aloud a short news-like printed passage. Although more constrained, this technique was considered appropriate for "assessing the mechanical skills of language production" (Underhill, 1987: 77).

A stimulus tape containing the 18 speech samples (six subjects performing three tasks) was put together. In order to curtail a carry over of one learner's rating on a certain task to a succeeding learner, an adaptation of the matched-guise technique (Lambert, 1967) was used to randomize the samples. As such, similar tasks or subjects were not placed in sequence on the stimulus tape.

2 Raters

The stimulus tape of the randomly ordered 18 speech samples was presented to 82 native speakers of Arabic for evaluation. The raters consisted of three groups: 1) 15 native speakers teaching AFL in the U.S. Analyses of data from the 15 subjects in this group were believed to yield stable results considering that the teachers were relatively homogeneous in terms of professional training; 2) 31 non-teaching native speakers of Arabic who had been residing in the U.S. for a period of at least one year. Raters within this group were all university students in central Ohio; and 3) 36 non-teaching native speakers of Arabic who were living in Lebanon. Raters within this group were university students.

3 Ratings

After listening to each speech sample, raters were given time--usually one-minute was sufficient--to provide their ratings. Anchored nine-point scales were used for the ratings, one indicating lowest performance level, and nine the educated native speaker. Each of the raters was requested to (1) provide a holistic score reflecting his/her overall impression of the level of proficiency of each of the 18 speech samples; and (2) provide ratings for each speech sample on specific unidimensional scales typically used in L2 oral assessment. The scales included

intelligibility, linguistic, and personality variables. Some of these scales were common across all three tasks and some were task specific.

III Research Design

In order to delineate the dimensions that raters considered when rating subjects' overall L2 oral language ability, multidimensional scaling (MDS) was employed. More specifically, individual differences scaling (INDSCAL) that "accounts for individual differences in the perceptual or cognitive processes that generate the responses [ratings]" (Young and Harris, 1990) was deemed as the most appropriate MDS technique to specify the salience of each of the delineated dimensions for each of the three rater groups.

The averaged holistic scores provided by each of the three rater groups were used to construct three proximity matrices, the rows and columns of which represented the 18 speech samples. The three matrices were submitted for analysis using the INDSCAL model within the ALSCAL MDS program of SPSS (1990).

It is important to note that the multidimensional solution was generated using dissimilarity matrices, which were based on holistic scores. The unidimensional scales were only used to assist in the interpretation of the ALSCAL solution.

IV Results

1 Dimensions of the Overall solution

ALSCAL solutions with two, three, and four dimensions were obtained. The nonmetric three-dimensional solution presented in this study was chosen on the basis of two criteria typically used to select the most appropriate dimensional representation of the data. The two criteria were: (1) fit indices; and (2) interpretability (Davison, 1983). The low average stress value of 0.116 and the high R^2 of 0.953 indicated that the nonmetric three-dimensional solution provided a good fit to the data. The three-dimensional space is presented in Appendix A.

In order to facilitate the interpretation of the three dimensions generated by the ALSCAL output, mean ratings on each of the unidimensional scales were regressed on the speech samples' stimuli coordinates. Because the resultant stimuli coordinates in MDS are orthogonal, i.e., the location of the stimuli on one dimension is independent of or uncorrelated with its location on the other dimension(s), the standardized regression coefficients can be thought of as correlations between the mean of the unidimensional rating scales and the stimuli location on each of the dimensions (Rocklin, 1992).

After performing regression analyses, two criteria were adopted to assist in selecting unidimensional scales that would best represent the dimensions in the ALSCAL solution. These two criteria involved regression weight patterns and meaningfulness of the scales in terms of speech sample analysis. The first criterion required that the magnitude of the dimension's regression weight on the selected unidimensional scale(s) be relatively high, and that the regression weights of the other dimensions on that scale(s) be low. The second criterion required that the selected unidimensional scales be meaningful and appropriate according to more qualitative speech sample analysis.

2 Dimension One

The first ALSCAL dimension to emerge was best defined by the two unidimensional scales "grammar" and "pronunciation" (see Table 1). The fact that both grammar and pronunciation represented dimension one was somewhat puzzling at first. Examining the speech samples closely, however, indicated how grammar and pronunciation can and do function jointly, and can be related because of inflectional markers (Alosh, unpublished manuscript). MSA has three short vowels /a/, /u/, and /i/ that are represented with diacritical marks placed above or below the letter they follow and function, mainly, as inflectional markers, i.e., case markers for nouns and mood markers for verbs. Therefore, the smallest inflectional mispronunciation is not only an error in pronunciation but also an error in grammar. The simple inflectional change can cause an error in verb tense, gender, etc. The dimension was, therefore, identified as grammar-pronunciation.

Table 1 Regressions Weights

Variable	Dim One	Dim Two	Dim Three
Grammar	-0.56	-0.26	-0.21
Pronunciation	-0.55	-0.38	-0.11
Creativity	-0.02	-0.76	-0.23
Adequacy of information	-0.21	-0.74	-0.07
Providing detail unassisted	-0.25	0.04	-0.82
Length of subject's responses	0.48	-0.22	-0.94

In order to investigate the interpretation of the grammar-pronunciation dimension, speech samples across the three tasks were examined. In the present paper, analysis of subject one's performance is reported. The location of subject one's stimuli along the grammar-pronunciation dimension, indicated that her performance on the interview had the highest dimensional value, then the narration, and lastly the read-aloud task. The magnitude of difference between her narration and interview locations was, however, small (see Appendix A). Examining her interview sample first, it was clear that subject one showed native-like pronunciation and made few grammatical errors. Her performance on the narration was not as good. Although her pronunciation of the sounds was still quite good, her speech was grammatically flawed. For example, subject one said "qarrara an yarkabu darrajatuhu," meaning the character decided to ride his bicycle. In order to illustrate the mistakes that demonstrate the association between grammar and pronunciation, the above sentence is broken down into three phrases:

	English	Arabic	Subject One	Correctness
1.	He decided	qarrara	qarrara	correct
2.	to ride	an yarkaba	an yarkabu	incorrect
3.	his bicycle	darrajatahu	darrajatuhu	incorrect

(he rides=yarkabu) (bicycle=darrajatu; his=hu)

The first phrase is correct. The second phrase is incorrect. In the second phrase, the word 'yarkabu' means 'he rides.' This is a present tense verb that takes the inflectional mark 'u' at the end of the word. When a present tense verb, however, is preceded by the particle

'an,' the inflectional mark 'u' changes to 'a.' The subject, however, failed to make the appropriate inflectional change and said 'an yarkabu' instead of 'an yarkaba.'

In the third phrase, the subject said 'darajatuhu,' meaning 'his bicycle.' When the word 'darrajatu' is in a nominal position it takes the nominal case marker 'u.' When, however, the word is in the accusative, as in the present example, then the ending changes to 'a,' and is pronounced 'darajatahu.' The subject, however, mispronounced 'darajatahu' as 'darajatuhu.' The above is one example of the type of mistakes that subject one made. This type of mistake in her speech explains the drop in her ranking on the grammar-pronunciation dimension.

Subject one's performance on the read-aloud task was not very different from the narration task. The subject continued to make mistakes similar to those encountered in the narration. To illustrate, the utterance "'limtu ...anna lisSan SaTa..." will be broken down similar to the previous example.

	English	Arabic	Subject One	Correctness
1.	I got to know	'limtu	'limtu	correct
2.	that	anna	an	incorrect
3.	a thief	lisSan	LisSan	correct
4.	broke into	SaTa	SaTan	incorrect

The first phrase is correct. Subject one mispronounced 'anna' as 'an,' in the second phrase. 'Anna' means 'that' and 'an,' as explained earlier, is a particle that precedes a present tense verb. The subject mispronounced 'SaTa' as 'SaTan,' which is a word that does not exist in Arabic. In short, subject one's rating dropped considerably on the narration and the read-aloud tasks because she made a number of grammatical mistakes.

3 Dimension Two

Dimension two was labeled creativity in presenting information. Dimension two was best represented by the unidimensional scales "creativity" and "adequacy of information in subject's narration" (see Table 1). The two unidimensional scales were part of the narration task. Upon examination of the speech samples, however, it became obvious that creativity in providing information was also a plausible dimension for the interview samples. Although it was only included in the narration unidimensional scales, raters were apparently using it when rating the interview speech samples. Creativity in presenting information in an interview context was considered in terms of interesting and engaging responses. Creativity in presenting information is not meaningful to the read-aloud task. The following discussion, therefore, will focus on the narration and interview tasks.

Analysis of subjects' speech samples corroborated the location of those samples on the dimension creativity in presenting information (see Appendix A). Subjects who were creative and engaging in presenting information scored highest on dimension two; whereas those who provided confusing information received lower scores. For example, subject two's narration included descriptions about the setting, and how the characters felt. These descriptions enlivened the story and provided the listener

with a feel for the story. Subject four's narration was without any embellishment that would capture the listener's attraction. As a result, subject two scored higher than subject four.

In contrast to the narration, subject four's performance on the interview, in terms of creativity in presenting information, was better than subject two's. This was also reflected in their location along dimension two where subject four received a higher dimensional value than subject two (see Appendix A). Analysis of the interview speech samples showed that subject four's interview was quite engaging. Her responses to the interviewer's questions were interesting and varied, and even humorous at times. For example, when asked whether she would like to get married, she responded saying "hopefully one day, although this is more difficult than studying." Subject two's interview was dry and more monotonous. He carried on about his father's profession saying "he used to work like a director. He was working in a place where he was going to take pictures of people and students and others also." In summary, both regression weights and analysis of speech samples indicated that creativity and ability to engage the listener meaningfully were deemed important to raters' judgements of the interview and narration tasks.

4 Dimension Three

Based on regression weights and speech sample analyses, dimension three was identified as amount of detail provided. Dimension three had the highest regression weights on the unidimensional scales "the ability of the subject to give detail unassisted" and "length of subject's responses" (see Table 1). Both of these scales were included in the interview unidimensional scales. Although the narration did not include these unidimensional scales, the narration speech sample analysis indicated that these two scales were meaningful. They reflected the amount of detail provided in subjects' narrations. With respect to the read-aloud task, giving detail was not applicable in a meaningful manner. The following discussion focused, therefore, on the narration and interview tasks.

Analysis of the interview speech samples verified the location of the stimuli on the amount of detail provided dimension. Subject one, for example, although one of the better speakers, she responded repeatedly with short answers and only after much probing on the part of the interviewer. To illustrate, subject one's performance on the interview was compared to the performance of subject three. The following is a sample of subject one's interview:

Interviewer: Have you visited a European country?
subject one: Yes, France.
Interviewer: How long did you stay in France?
subject one: Two months.
Interviewer: In what city?
subject one: Toulouz."
Interviewer: Is it pretty?
subject one: Yes, very pretty.
Interviewer: Is it prettier than Paris?
subject one: No, Paris is prettier.

As evidenced here, although subject one was providing responses to the questions, she was not volunteering any additional information. She did not elaborate or explain what she meant. Subject three, on the other hand, described with detail why he liked Cairo, (busy streets, hot weather, great pyramids...). When asked whether he worked while studying, he provided his work schedule, and described his duties at the restaurant and the adolescence center. The difference in the amount of detail provided by each of these two subjects explained their respective locations on dimension three, i.e., subject one received lower scores than subject three (see Appendix A).

With respect to the narration, examining the amount of detail provided in subject three's and subject one's narrations, again, explicated their respective rankings on the dimension amount of information provided, i.e., subject one received a lower score than subject three (see Appendix A). In narrating, subject three discussed every picture individually, stating the number of the picture and describing it. Subject one, on the other hand, presented a more summarized account of the visuals. For example, in describing the accident, subject one said "the young man saw the beautiful girl and did not see the road and he broke his leg." Subject three, however, said "the man looked at a girl wearing short pants and he got into an accident with a big tree and thus got off the road...the man has a broken leg and arm." In short, subject one, similar to her performance on the interview, was very brief and provided no detail. Subject three, however, elaborated and provided more detail.

5 Individual Differences

In addition to the multidimensional solution, the INDSCAL model on ALSCAL computes individual weight matrices. The weight matrices account for individual differences among raters in the importance or salience of each of the dimensions in the solution.

Previous research has indicated that teachers tend to emphasize grammar in their assessment of students' proficiency and non-teachers tend to be concerned with the more communicative aspects of the language. According to the literature, therefore, we would expect the teaching group to emphasize dimension one, i.e., the grammar-pronunciation dimension, and the non-teaching groups to rely more on the other two dimensions--creativity in providing information and amount of detail provided.

In the present study, INDSCAL provided weights that depicted the extent to which each of the three rater groups relied on each of the dimensions in their holistic rating of subjects' L2 oral ability. Subject weights presented in Table 2 indicate that the three rater groups were emphasizing different criteria in judging subjects' overall performance. The group of non-teaching Arabs in the U.S. emphasized all three dimensions in their ratings, although dimension three, quantity of speech, had the most salience. Raters in the teaching group seemed to be relying most heavily on dimension two, creativity in presenting information. The group of non-teaching Arabs residing in Lebanon, however, emphasized almost solely the grammar-pronunciation dimension, i.e., dimension one.

Table 2 Rater group Weights

Rater Group	Dimension One	Two	Three
U.S. residents	0.4917	0.3470	0.7642
Teachers	0.0061	0.9352	0.2922
Lebanese	0.9752	0.0087	0.0395

Results of the present study were not consistent with research (Chastain, 1980) indicating that teachers tend to emphasize grammar in their assessment of students' proficiency and non-teachers tend to be concerned with the more communicative aspects of the language. Results reported in the literature were based, however, on studies using languages other than MSA. More specifically, results in the present study could have been due to diglossia in Arabic.

In a diglossic situation, two varieties of the same language exist side-by-side, each variety having a specialized function. Two forms of the Arabic language co-exist in the Arab World: MSA and the colloquial variations of Arabic. MSA is a written and a spoken language used for formal instruction, general lectures, official correspondence, administrative announcements, and in mass media. MSA is readily understood by educated Arabs. Colloquial Arabic is used for everyday activities. It comprises the local spoken dialects, which are acquired natively by Arabs, and is not readily understood by all Arabs.

Because MSA is the form used in most AFL classes and because AFL classes increasingly emphasize communicative language teaching and learning, the teaching rater group are probably accustomed to using MSA not only in its typical domains but also for everyday activities, which may explain why teachers emphasized the more communicative aspects when rating the subjects.

The non-teaching Lebanese group, however, probably use MSA in its more formal contexts where accuracy plays a central role. Thus in judging the L2 oral ability of the subjects in the present study, this group relied to a large extent on the grammar-pronunciation dimension.

Two factors may explain the reasons why the group of non-teaching raters in the U.S. differed from the non-teaching Lebanese group: (1) those in the U.S. may be aware of the communicative-based AFL classroom situation; and/or (2) they have been increasingly using MSA to communicate with other Arabs around them, because, as mentioned earlier, MSA is readily understood by educated Arabs.

V Conclusion

L2 research has documented variability in language performance across different tasks. This phenomenon was also evidenced in the present study. When explicating subject one's performance along the grammar-pronunciation dimension, we saw that her performance varied from task to task. Variability across tasks was also evidenced in subject two's and subject four's performance in terms of creativity in presenting information on the narration and interview tasks. Whereas subject two

outperformed subject four on the narration, subject two's performance had lower dimensional value than subject four's on the interview.

Performance variability may be attributed to some extent to the difference in the demands that the task places on the linguistic and cognitive processes of the subjects thus influencing their performance. For example, in the interview, the interviewer was present to interact with the subjects and to direct their efforts in constructing their speech. In the read-aloud, subjects were provided with a text that obviously constrained their language production. Also, unlike the interview, the read-aloud did not allow for interaction with another speaker and for immediate feedback. In the narration, subjects were not as constrained with a selected text or set of questions. They were required, however, to interpret and present the visuals without access to any linguistic support or feedback.

Variability, however, was not evident on all three dimensions. Whereas subjects' performances varied somewhat from task to task when examined in terms of the grammar-pronunciation and creativity in presenting information dimensions, subjects' performance, remained relatively unchanged when analyzed in terms of the third dimension, amount of detail provided. In summary, some aspects in learners' performance may be relatively stable from task to task and others may vary somewhat. More research is needed that investigates those dimensions in L2 oral production that are stable across tasks in contrast to those that vary.

In addition to variation in subjects' language products due to task, rater groups vary in their expectations and evaluations from task to task. The teaching rater group considered those dimensions that are more in tune with the profession's shift in focus to more communicative assessment.

The non-teaching rater groups, who do not have professional training, were operating with a different set of criteria. As hypothesized, the Lebanese rater group differed from those residing in the U.S. The U.S. rater groups were more diversified in terms of the dimensions they employed when rating subjects' L2 oral performance while the Lebanese seemed to emphasize the grammar-pronunciation dimension.

Who should, therefore, be used as the rater criterion in L2 oral assessment? Should different criteria be adopted for different purposes? Should the criteria that teachers emphasize be upheld, for example, when academic pursuit is the goal? On the other hand, should subjects be tested for those aspects of the language that non-teaching native speakers deem important when the goal is interacting with non-teaching native speakers? Further research needs to investigate these issues as they influence the use and interpretation of L2 oral scores.

In summary, besides L2 oral ability, both tasks and raters affect students' L2 oral scores. Tasks place different demands on students and alters their performance, while raters emphasize different aspects in their ratings. Therefore, both tasks and raters warrant attention as they influence the validity of the L2 oral construct. L2 oral testing researchers might need reconsider employing generic, component scales.

It is recommended that scales be empirically derived according to the given tasks and audiences.

In the context of the present paper, further analyses need to be performed on each of the three tasks separately. It would be interesting to observe whether the dimension that emerged when investigating all three tasks together would still come out when analyses are performed on each of the three tasks separately. It would also be interesting to investigate whether the same pattern of subject weights would be generated. Finally, it is important to note that findings reported in the present paper need to be validated with other languages, oral ability levels, tasks, and rater groups.

VI References

Alosh, M. 1988: Underrepresentation in Arabic: A comprehension aid or hinderance? Unpublished manuscript.

Bachman, L.F. 1990: Fundamental considerations in language testing. Oxford: Oxford University Press.

Bachman, L.F., and Palmer, A.S. 1981: A multitrait-multimethod investigation into the construct validity for six tests of speaking and reading. In Palmer, A.S., Groot, P.J.M. and Trosper, G.A., editors, The construct validation of tests of communicative competence, Washington, D.C.: Teachers of English to Speakers of Other Languages, 149-65.

Barnwell, D. 1989: Naive' native speakers and judgements of oral proficiency in Spanish. Language Testing 6, 152-63.

Chastain, K. 1980: Native speaker reaction to instructor-identified student second-language errors. Modern Language Journal 64, 210-15.

Clifford, R.T. 1981: Convergent and discriminant validation of integrated and unitary language skills: The need for a research model. In Palmer, A.S., Groot, P.J.M. and Trosper, G.A., editors, The construct validation of tests of communicative competence, Washington, D.C.: Teachers of English to Speakers of Other Languages, 149-65.

Davison, M. 1983: Introduction to multidimensional scaling and its applications. Applied Psychological Measurement. 373-79.

Ellis, R. 1985: Understanding second language acquisition. Oxford: Oxford University Press.

Ellis, R. 1987: Second language acquisition in context. Englewood Cliffs, NJ: Prentice-Hall International.

Engber, C. 1987: Summary of the discussion session. In Valdman, A., editor, Proceedings of the symposium on the evaluation of foreign language proficiency. Bloomington, Indiana: Committee for Research and Development in Language Instruction.

Galloway, V.B. 1980: Perceptions of the communicative efforts of American students of Spanish. Modern Language Journal 64, 428-33.

Lambert, W.E. 1967: The social psychology of bilingualism. Journal of Social Issues 23, 91-109.

Larsen-Freeman, D. and Long, M. 1991: An introduction to second language acquisition research. New York: Longman.

Messick, S. 1993: Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In Bennett, R.E. and Ward, W.C., editors, Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. Hillsdale, NJ: Lawrence Erlbaum Associates, 61-73.

Moss, P.A. 1992: Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research 62, 229-58.

Omaggio, A.C. 1986: Teaching language in context: Proficiency-oriented instruction. Boston, MA: Heinle & Heinle Publishers, Inc.

Rocklin, T. 1992: A multidimensional scaling study of college students' perceptions of test item formats. Applied Measurement in Education 5, 123-36.

Shohamy, E. 1983: The stability of oral proficiency assessment on the oral interview testing procedure. Language Learning 33, 527-39.

Shohamy, E. 1984: Does the testing method make a difference? The case of reading comprehension. Language Testing 1, 147-70.

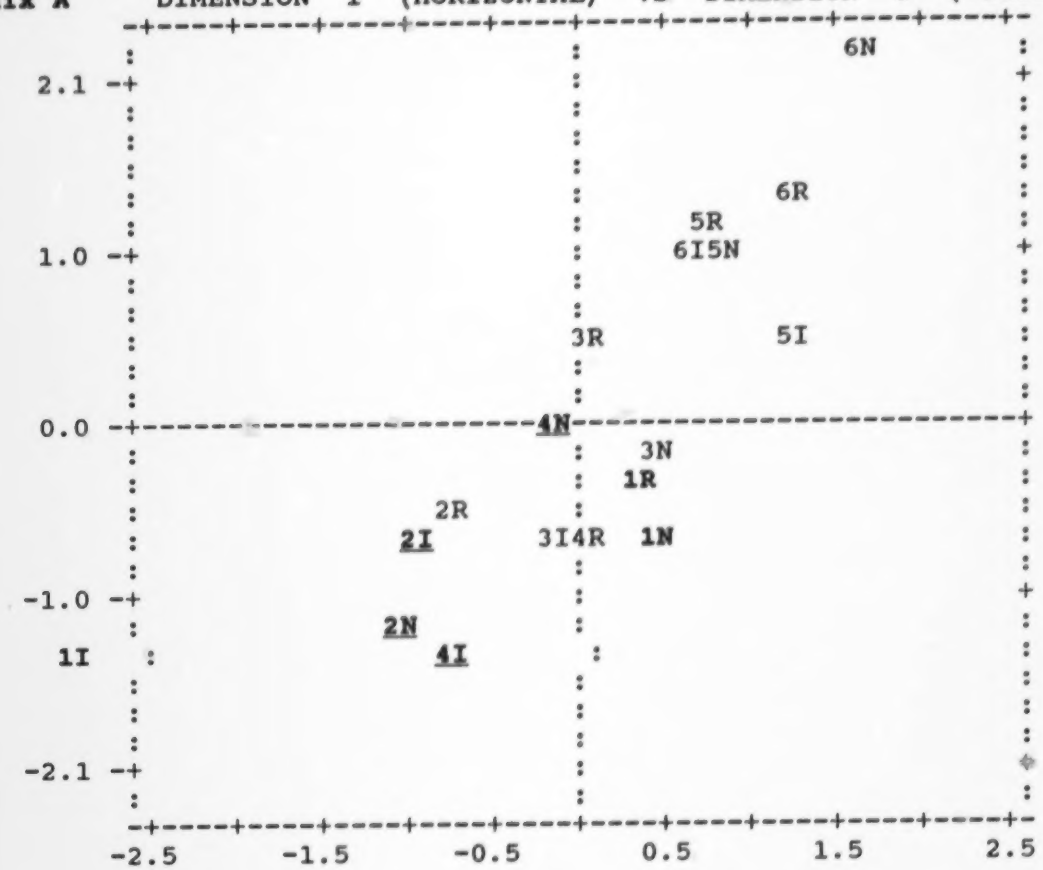
Shohamy, E., Gordon, C.M. and Kraemer, R. 1992: The effect of raters' background and training on the reliability of direct writing tests. Modern Language Journal 76, 27-33.

Tarone, E. 1989: Accounting for style-shifting in interlanguage. In Gass, S., Madden, C., Preston, D. and Selinker, L. editors, Variation in second language acquisition volume II: Psycholinguistic issues 13-21.

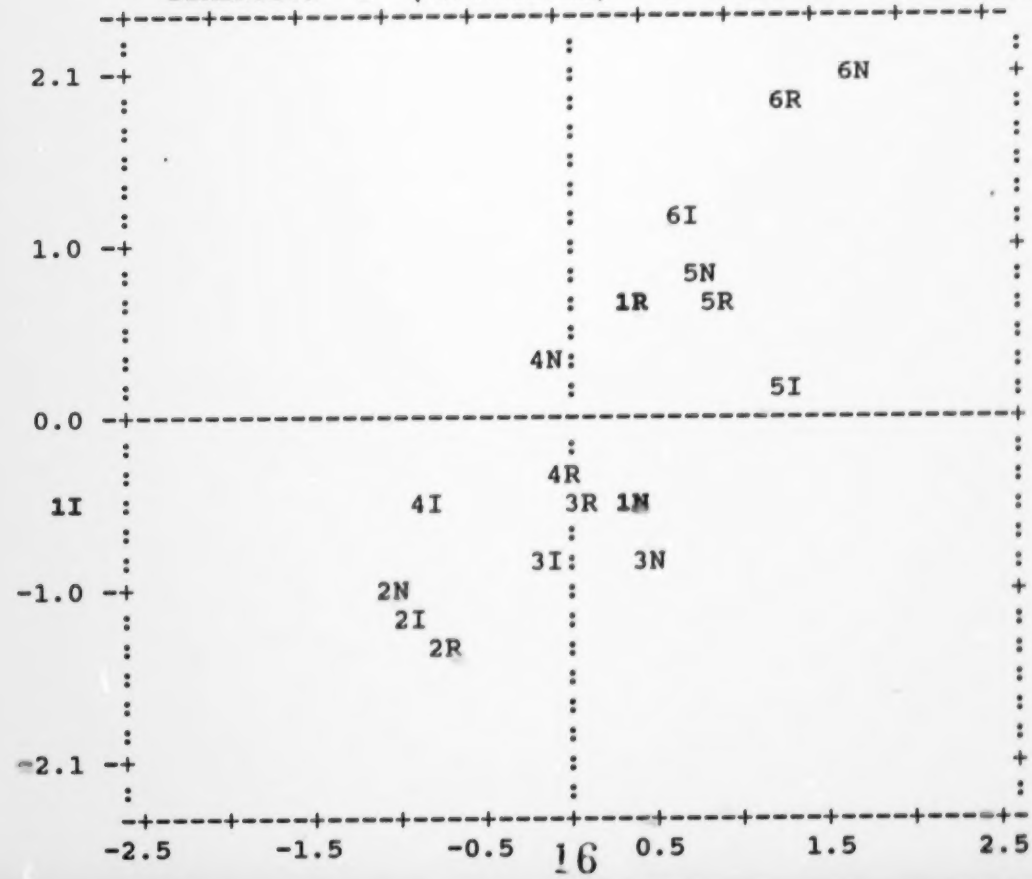
Underhill, N. 1987: Testing Spoken Language: A handbook of oral testing techniques. New York: Cambridge University Press.

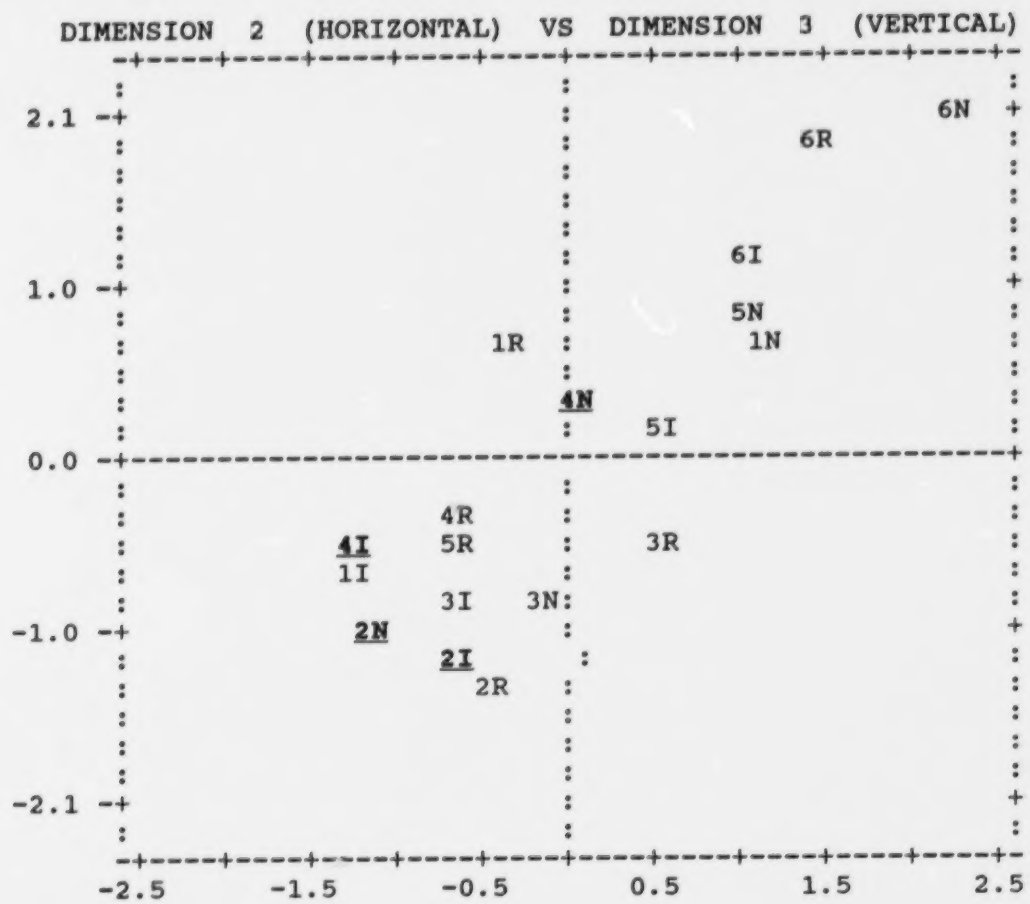
Young, F.W. and Harris, D.F. 1990: Multidimensional scaling: Procedure ALSCAL. In SPSS user's guide. Chicago, IL: SPSS Inc.

Appendix A DIMENSION 1 (HORIZONTAL) VS DIMENSION 2 (VERTICAL)



DIMENSION 1 (HORIZONTAL) VS DIMENSION 3 (VERTICAL)





Stimulus coordinates

Subject Number	Task	Dimension One	Two	Three
1	interview	-2.7195	-1.3321	-0.6265
3	narration	0.4162	-0.1890	-0.8231
1	read-aloud	0.3488	-0.4242	0.6128
2	narration	-1.0726	-1.1865	-1.0055
3	interview	-0.1833	-0.7407	-0.8921
4	narration	-0.1935	-0.0187	0.3845
5	interview	1.1935	0.4986	0.1926
6	narration	1.5620	2.2126	2.0428
3	read-aloud	-0.0488	0.5425	-0.6050
6	interview	0.5951	1.0199	1.2956
2	read-aloud	-0.7881	-0.4951	-1.3251
4	interview	-0.9086	-1.3202	-0.5692
6	read-aloud	1.1601	1.3881	1.8205
5	narration	0.7273	1.0481	0.8269
2	interview	-0.9881	-0.6638	-1.2442
4	read-aloud	-0.0999	-0.7474	-0.2603
1	narration	0.3263	-0.6559	-0.4762
5	read-aloud	0.6731	1.0636	0.6515